Jörg Tiedemann
Department of Digital Humanities
University of Helsinki

Found in Translation 2024

Natural Language Understanding with Multilingual Data

An Analysis of Encoder Representations in Transformer-Based Machine Translation

Alessandro Raganato and Jörg Tiedemann
Department of Digital Humanities
University of H
{alessandro.raganato, jorg.

Abstract

Fixed Encoder Self-Attention Patterns in Transformer-Based Machine Translation

Alessandro Raganato, Yves Scherrer and Jörg Tiedemann
University of Helsinki
}@helsinki.fi

2019; Voita et al., 2019a; Brunner et al
A closely related research area attempts
the attention mechanism, e.g. by inco

On the differences between BERT and MT encoder spaces and how to address them in translation tasks

Raúl Vázquez Hande Celikkanat Mathias Creutz Jörg Tiedemann
Department of Digital Humanities
University of Helsinki

Tracking the Traces of Passivization and Negation in Contextualized Representations

Hande Celikkanat Sami Virpioja Jörg Tiedemann Marianna Apidianaki
Department of Digital Humanities
University of Helsinki

A Closer Look at Parameter Contributions When Training Neural Language and Translation Models

Raúl Vázquez[♠] Hande Celikkanat[♠] Vinit Ravishankar[◇] Mathias Creutz[♠] Jörg Tiedemann[♠]
[♠]Department of Digital Humanities, University of Helsinki
[◇]Language Technology Group, Department of Informatics, University of Oslo
[♠]{firstname.lastname@}@helsinki.fi [◇]vinitr@ifi.uio.no

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 115 OCTOBER 2020 143–162

Are Multilingual Neural Machine Translation Models Better at Capturing Linguistic Features?

David Mareček,[a] Hande Celikkanat,[b] Miikka Silfverberg,[b] Vinit Ravishankar,[c] Jörg Tied

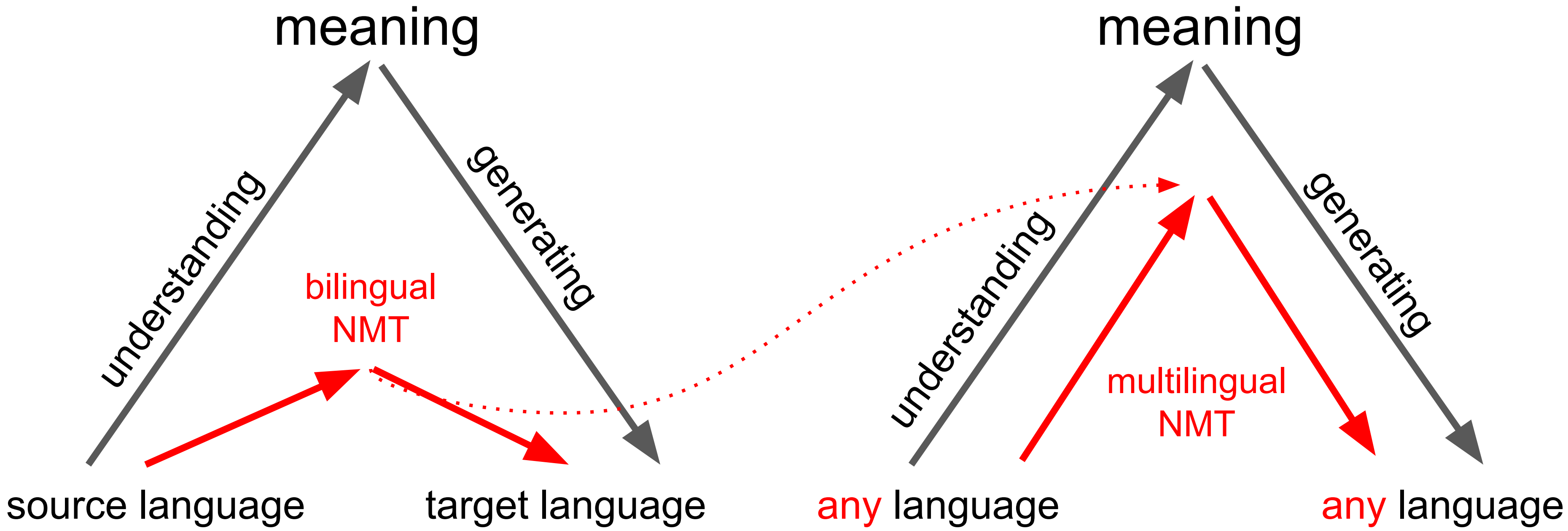
The idea: Use translations to learn representations

visual grounding

En:	A *wall* divided the city.	
De 1:	Eine *Wand* teilte die Stadt.	×
De 2:	Eine *Mauer* teilte die Stadt.	✓

“translational grounding”

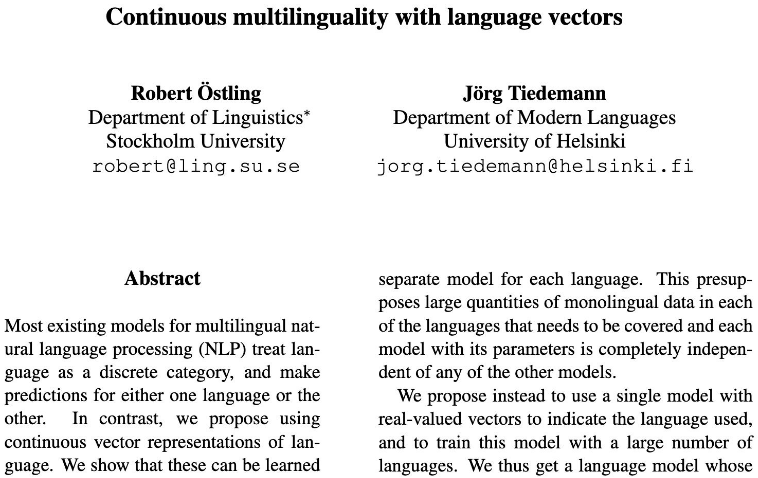
The hypothesis: Linguistic diversity helps



A starting point: A character-LM for ca. 1000 languages

Back in 2016:

1303 Bible translations
into 990 languages



Continuous multilinguality with language vectors

Robert Östling
Department of Linguistics*
Stockholm University
robert@ling.su.se

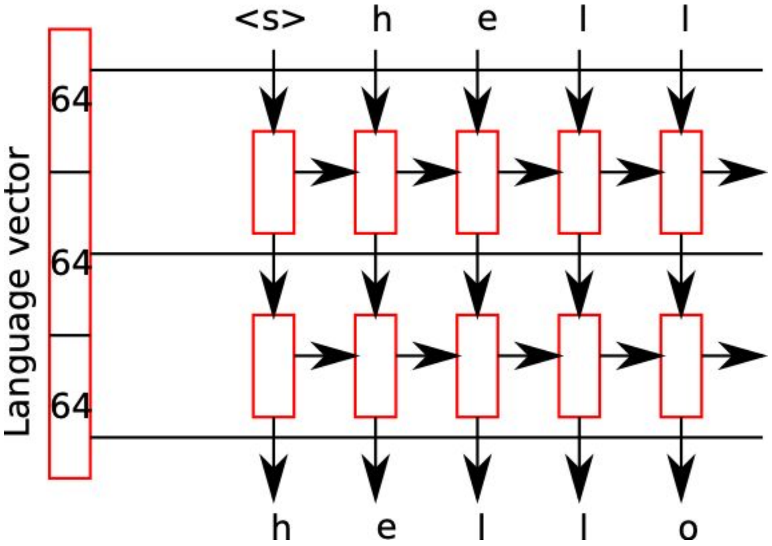
Jörg Tiedemann
Department of Modern Languages
University of Helsinki
jorg.tiedemann@helsinki.fi

Abstract

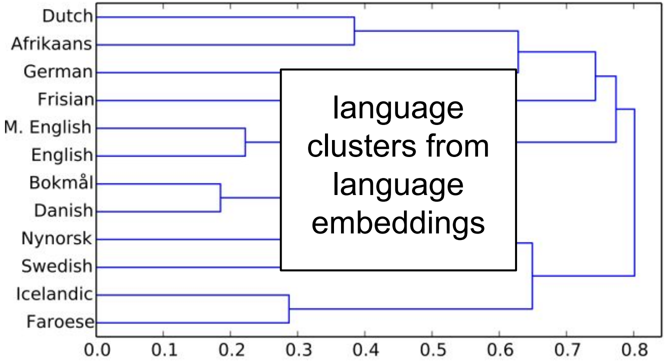
Most existing models for multilingual natural language processing (NLP) treat language as a discrete category, and make predictions for either one language or the other. In contrast, we propose using continuous vector representations of language. We show that these can be learned

separate model for each language. This presupposes large quantities of monolingual data in each of the languages that needs to be covered and each model with its parameters is completely independent of any of the other models.

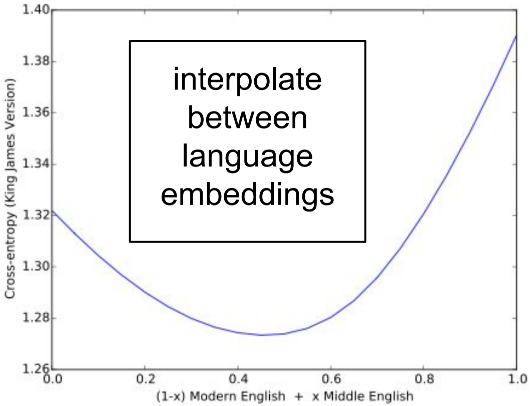
We propose instead to use a single model with real-valued vectors to indicate the language used, and to train this model with a large number of languages. We thus get a language model whose



The language continuum and language embeddings



language clusters from language embeddings



interpolate between language embeddings

middle English

modern English

Control text generation with language embeddings:

turn on Swedish:

och jehova sade till honom : ” jehova har sagt , och jag skall ...

turn on German:

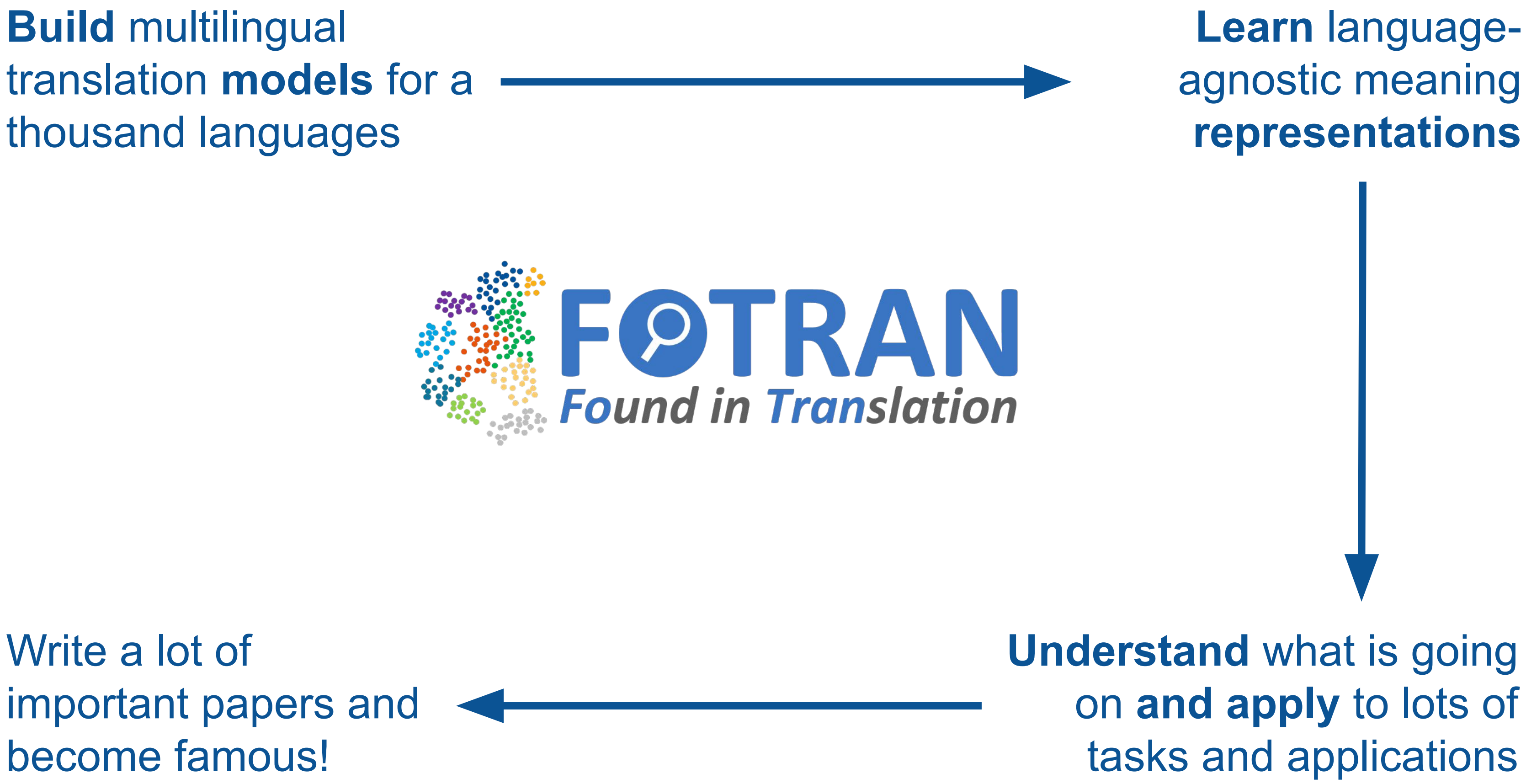
und er sprach zu ihnen : siehe , ich bin der herr

mix Swedish and German:

vocken ånner vocken ånnen söhenöckenföcken ...

average of Scandinavian languages:

og han sa til herrens : " han skal vitnaðus til herrens hjárt



Build multilingual translation **models** for a thousand languages

Learn language-agnostic meaning **representations**

Understand what is going on **and apply** to lots of tasks and applications

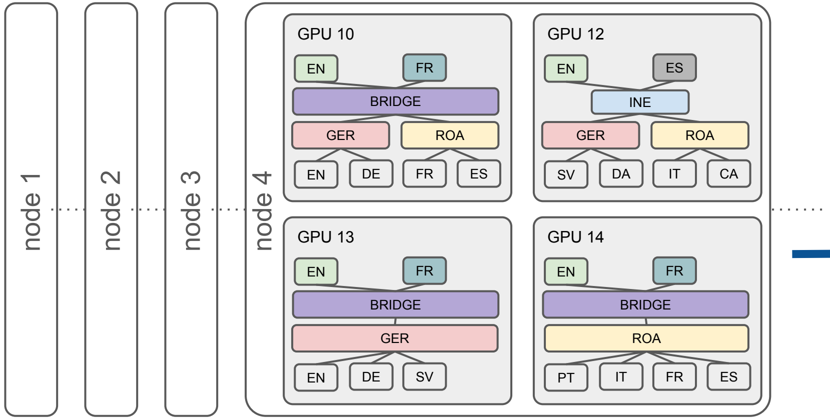
Write a lot of important papers and become famous!

FOTRAN

Found in Translation

FoTran in a nutshell

Building large multilingual neural translation models



node 1

node 2

node 3

node 4

GPU 10

EN FR

BRIDGE

GER ROA

EN DE FR ES

GPU 12

EN ES

INE

GER ROA

SV DA IT CA

GPU 13

EN FR

BRIDGE

GER

EN DE SV

GPU 14

EN FR

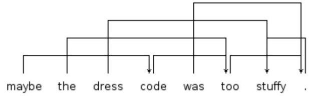
BRIDGE

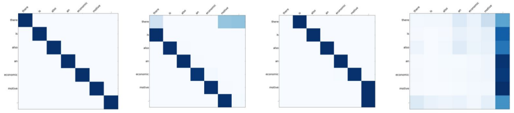
ROA

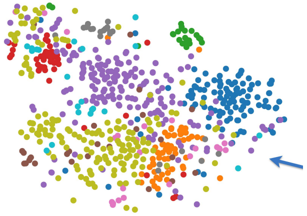
PT IT FR ES

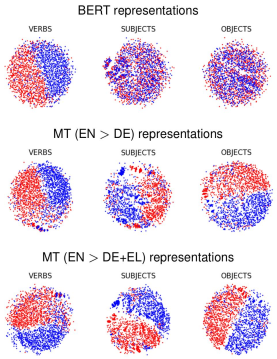
Creating and evaluating downstream applications such as machine translation

Interpretability and analysis









FOTRAN
Found in Translation

erc

Proposal written, submitted
and … even accepted

FoTran starts: language
embeddings, analysing transformers

NLI, Modular NMT &
the attention bridge model

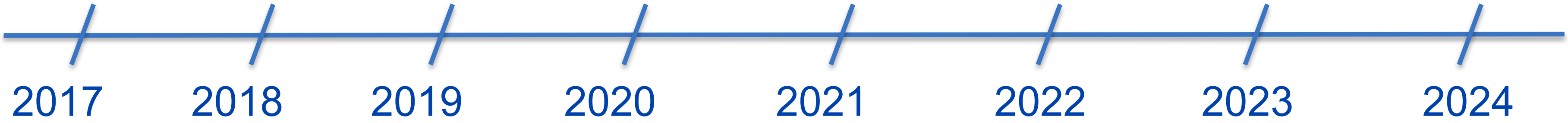
OPUS-MT, Tatoeba challenge
analyse contextualized embeddings

Fixed attention patterns
LMs vs MT models

NLU benchmark diagnostics
LM and MT training dynamics

Scaling-up modular models
MAMMOTH

Goodbye!



LASER

OpenAI
GPT-3

NLLB

SONAR
SeamlessM4T

Jörg Tiedemann
Department of Digital Humanities
University of Helsinki

Found in Translation 2024

Goodbye FoTran!

An Analysis of Encoder Representations in Transformer-Based Machine Translation

Alessandro Raganato and Jörg Tiedemann
Department of Digital Humanities
University of H
{alessandro.raganato, jorg.

Abstract

Fixed Encoder Self-Attention Patterns in Transformer-Based Machine Translation

Alessandro Raganato, Yves Scherrer and Jörg Tiedemann
University of Helsinki
}@helsinki.fi

2019; Voita et al., 2019a; Brunner et al
A closely related research area attempts
the attention mechanism, e.g. by inco

On the differences between BERT and MT encoder spaces and how to address them in translation tasks

Raúl Vázquez Hande Celikkanat Mathias Creutz Jörg Tiedemann
Department of Digital Humanities
University of Helsinki

Tracking the Traces of Passivization and Negation in Contextualized Representations

Hande Celikkanat Sami Virpioja Jörg Tiedemann Marianna Apidianaki
Department of Digital Humanities
University of Helsinki

A Closer Look at Parameter Contributions When Training Neural Language and Translation Models

Raúl Vázquez♠ Hande Celikkanat♠ Vinit Ravishankar◇ Mathias Creutz♠ Jörg Tiedemann♠
♠Department of Digital Humanities, University of Helsinki
◇Language Technology Group, Department of Informatics, University of Oslo
♠{firstname.lastname@}@helsinki.fi ◇vinitr@ifi.uio.no

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 115 OCTOBER 2020 143–162

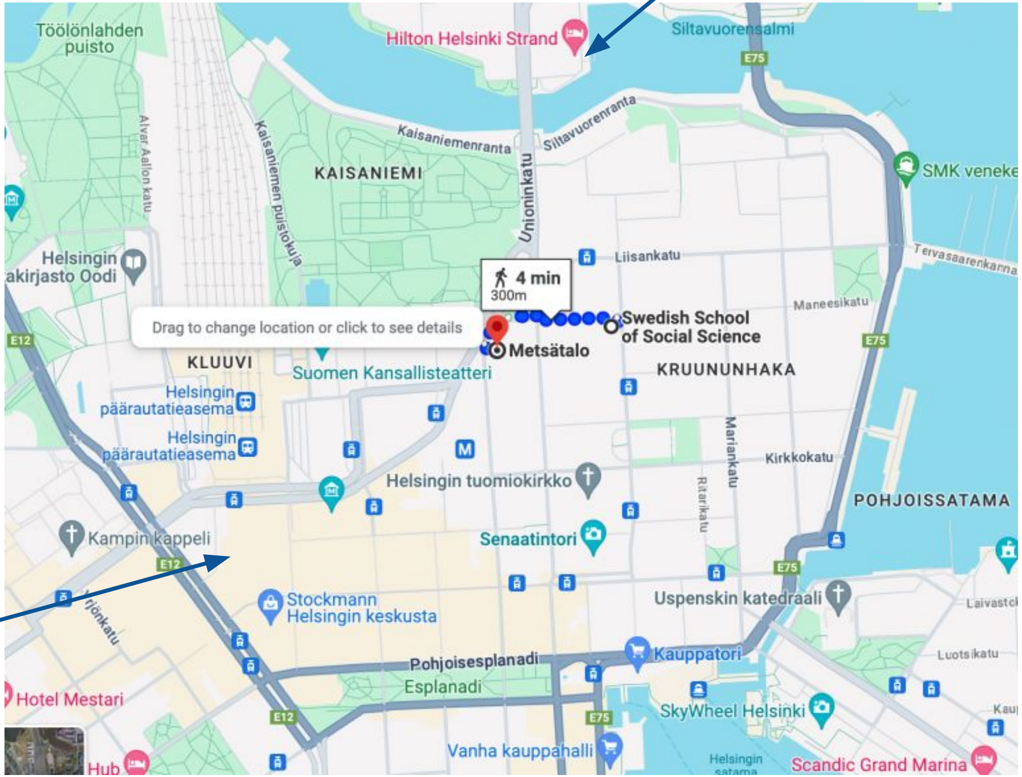
Are Multilingual Neural Machine Translation Models Better at Capturing Linguistic Features?

David Mareček,[a] Hande Celikkanat,[b] Miikka Silfverberg,[b] Vinit Ravishankar,[c] Jörg Tied

[a] Institute of Formal and A

The logistics

- **Date:** Thursday, February 22, 2024
- **Place:** University of Helsinki, Central Campus
    - morning session: Soc & Kom, room 210, Snellmaninkatu 12, Helsinki
    - afternoon session: Metsätalo, room B214 (hall 4), Unioninkatu 40, Helsinki



Lunch:
Restaurant Bro

Dinner:
Restaurant Zetor

https://blogs.helsinki.fi/language-technology/goodbye-fotran/

The Program

Morning coffee

- 10:00 – Welcome and a short background on the FoTran project
- 10:30 – Alessandro Raganato (University of Milano-Bicocca)
- 11:15 – Marianna Apidianaki (University of Pennsylvania)

Lunch Break

- 14:00 – Vered Shwartz (University of British Columbia)
- 15:30 – Poster/demo session with snacks and refreshments

Dinner

Tomorrow, Feb 23: FoTran PhD Defence – Aarne Talman

- **Place:** Room 303, Unioninkatu 33, Helsinki

https://blogs.helsinki.fi/language-technology/goodbye-fotran/